Proceedings Paper:
Tzagarakis, G. and Panoutsos, G. (2016) Model-Based Feature Selection Based on Radial Basis Functions and Information Measures. In: 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2016 IEEE World Congress on Computational Intelligence (WCCI), 24-29 July 2016, Vancouver, BC, Canada. Institute of Electrical and Electronics Engineers (IEEE) , pp. 401-407. ISBN 9781509006267

https://doi.org/10.1109/FUZZ-IEEE.2016.7737715

Model-Based Feature Selection Based on Radial Basis Functions and Information Measures

Georgios N. Tzagarakis
Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield, UK
g.tzagkarakis@sheffield.ac.uk

George Panoutsos
Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield, UK
g.panoutsos@sheffield.ac.uk

***Abstract*—In this paper the development of a new embedded feature selection method is presented, based on a Radial-Basis-Function Neural-Fuzzy modelling structure. The proposed method is created to find the importance of features in a given dataset (or process in general), with special focus on manufacturing processes. The proposed approach evaluates the impact/importance of processes features by using information theoretic measures to measure the correlation between the process features and the modelling performance. Crucially, the proposed method acts during the training of the process model; hence it is an embedded method, achieving the modelling/classification task in parallel to the feature selection task. The latter is achieved by taking advantage of the information in the output layer of the Neural Fuzzy structure; in the presented case this is a TSK-type polynomial function. Two information measures are evaluated in this work, both based on information entropy: mutual information, and cross-sample entropy. The proposed methodology is tested against two popular datasets in the literature (IRIS – plant data, AirFoil – manufacturing/design data), and one more case study relevant to manufacturing – the heat treatment of steel. Results show the good and reliable performance of the developed modelling structure, on par with existing published work, as well as the good performance of the feature selection task in terms of correctly identifying important process features. In the presented case studies and simulation results the mutual-information – based implementation of the algorithm appears to perform better compared to the cross-sample entropy-based implementation.**

Keywords: Feature selection, information entropy, information measures, Radial Basis Function, Fuzzy Logic, Manufacturing Systems

I. INTRODUCTION

Feature selection is the procedure for finding the most important features of a system by removing irrelevant data (or whole variables). This is often a significant step in data-driven modelling, in order to develop models that represent the behaviour of the process under investigation. The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining high accuracy in representing the original features. Good feature subset includes features that are correlated with the decision feature and uncorrelated with each other.

Feature selection algorithms are categorised into filters, wrappers and embedded methods [1, 2]. In this paper, an embedded method is presented; this is a method that combines the construction of the classifier/model and the feature selection task. The proposed method relies on a Radial Basis Function (RBF) classifier (one that is designed to be equivalent to a Fuzzy Logic–based system) to perform the classification task, while at the same time also performs a ranked feature selection. For the first time in the literature, information measures are utilised to perform the embedded feature selection in an RBF system, namely: mutual information and Cross-Sample Entropy. These information measures are used to measure the relevance of the individual features compared to the performance of the model (prediction accuracy). For the RBF implementation a 3-layer Neural Network is used, with the output layer defined as the equivalent of the TSK Fuzzy Logic System. Subsequently, information measures are applied to the output layer, to perform feature selection.
Mutual Information (MI) [3], Approximate Entropy (ApEn) [4] and Cross-Sample Entropy (CSE) [5] are some of the most commonly used algorithms for feature selection that are based on Shannon entropy [6].

Several studies focus on the use of mutual information and cross-sample entropy, as filter methods, to perform the feature selection task. For example, in the area of healthcare, CSE is used in RNA structure analysis [7] and DNA microarray analysis [8], as well as MI has been used in [3] for feature selection. The use of such information measures is also popular with wrapper (use of classifiers) such Naïve Bayes [9], Support Vector Machine [10], Probabilistic Neural-Network [11] as well as clustering methods: k-nearest neighbour [12] and Decision Trees [13]. Existing work uses entropy methods to pre- or post- process the results (raw datasets) of wrapper (classifiers) methods. There is existing work that addresses Fuzzy Logic and wrapper- or embedded-based feature selection, such as [14-16], however no work has been so far reported that focuses on information measures and RBF Fuzzy Logic Systems as an embedded method, i.e. a method that performs classification and feature selection in one task.

The presented work relies on a popular implementation of Fuzzy Logic systems, the Radial-Basis-Function Neural-Network [17]. This implementation, as shown in detail in the following sections conveniently uses a simple 3-layer Neural-Network (NN) structure to realise a Fuzzy Logic equivalent modelling structure, under some conditions. The output layer (in this case a TSK polynomial function) of the NN is used to extract information on the relevance of the process features to the performance of the model. This is carried out while the NN is trained via an error-propagation (EP) parametric optimisation routine. Thus, the proposed embedded feature selection method performs the feature selection task while the NN is trained (by utilising information produced after each iteration of the EP algorithm. The proposed work is tested against publicly available benchmark data, as well as a real case study on a manufacturing process that is highly non-linear and contains significant uncertainty in the data. Results show that the proposed algorithm performs well, and correctly identifies the relevant features in every case, while also achieving a very good classification performance.

The rest of this paper is organised as follows: a brief description of each of the used algorithms and computational methods is presented in Section II: Background Theory. The proposed method is presented in Section III, and associated simulation results are shown in Section IV. Finally, Section V includes concluding remarks on the proposed feature selection approach and directions for future research.

II. Background Theory

A. Radial Basis Function Neural-Fuzzy Modelling

Neural-Fuzzy models are popular implementations of Fuzzy Logic Systems due to their hybrid modelling characteristics, which share traits from Neural-Networks as well as Fuzzy Logic Systems [18]. Specifically, the learning performance of the NN, is combined with the transparency, simplicity and tolerance to uncertainty of the Fuzzy Logic system. Radial Basis Functions can be used as the activation functions of a simple 3-layer Neural-Network to create a modelling structure that is mathematically equivalent to a Fuzzy Logic system [19]. Fig. **1** shows the structure of the RBF-NF model, which is mathematically described as:

$$y=\sum_{i=1}^{p} z_i \left[\frac{\prod_{j=1}^{m} \mu_{ij}\left(x_j\right)}{\sum_{i=1}^{p}\prod_{j=1}^{m} \mu_{ij}\left(x_j\right)} \right] \qquad (1)$$

where $\mu_{ij}(\chi_j)$ is the Gaussian membership function of χ_j that belongs to the *i-th* rule.

$$\mu_{ij}\left(x_j\right)=e^{\left(-\frac{\left(x_j-c_{ij}\right)^2}{\sigma_{ij}^2}\right)} \qquad (2)$$

where c_{ij} and σ_{ij} are the centre and the width of each membership function respectively, m the number of inputs and p number of rules.

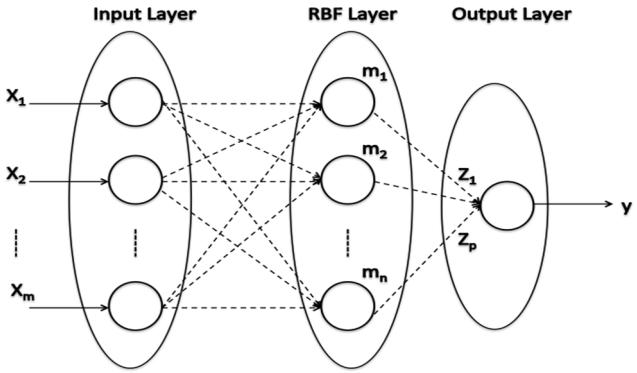

Fig. 1 Radial Basis Function – Neural Fuzzy Model

This NFM implementation is for centre of gravity defuzzification, product inference rule and an output function of singleton type. Mamdani and TSK implementations are also possible by replacing the output layer of the NN with appropriate functions.

The output of the Neural-Network can be calculated using the definition of the RBF function as follows:

Radial Basis Function:

$$g_i\left(x\right)=\frac{m_i\left(x\right)}{\sum_{i=1}^{p} m_i\left(x\right)} \qquad (3)$$

Neural-Fuzzy-Model

$$y=\frac{\sum_{i=1}^{p} z_i m_i\left(\underline{x}\right)}{\sum_{i=1}^{p} m_i\left(\underline{x}\right)} \qquad (4)$$

where $m_i\left(\underline{x}\right)=e^{\left(-\left\|\underline{x}-\underline{c_i}\right\|^2/\underline{\sigma_i^2}\right)}$ is the membership degree of the input vector.

The training of the RBF-NFM can be achieved by a number of parametric optimisation algorithms, such as ones based on gradient descent or evolutionary optimisation methods. In this paper an error propagation (EP) algorithmic used, with adaptive learning and momentum rates for better avoidance of local minima [19].

The appeal of the RBF-NFM is that the overall model structure is rather simple (3-layer NN), thus computationally not expensive, it offers universal approximation capability [20], and it is mathematically equivalent to a class of Fuzzy Logic systems.

B. Information Measures

Mutual Information (MI). One of the goals in predictive modelling is to minimise the uncertainty of the dependent variable. A good formalisation of the uncertainty of a random variable is given in Shannon and Weaver's [21]. MI is a criterion from the information theory and has proven very efficient feature selection algorithm [22, 23]. The mutual information measures the amount of information contained in a variable or a group of variables, in order to predict the dependent one. It also is model-independent, and nonlinear, as it measures the nonlinear relationships between variables. One of the most important advantages of MI is its ability to detect non-linear relationships between variables, while other popular criteria as the well-known correlation coefficient are limited to linear relationships. Using MI it is possible to process both categorical and discrete data [24].

Cross-Sample Entropy. The Cross-Sample Entropy (CSE) is an extension of the sample entropy (Samp-En) algorithm and was introduced by Richman and Moorman [25]. Samp-En is also introduced by Richman and Moorman [25] and has the ability to estimate the signals' regularity. One has to divide the time-series in subseries with length *m* and estimate the conditional probability of how times matches to the next subseries with the same length m and with a tolerance *r*. The negative natural algorithm of the previous result is the Samp-En. Samp-En is introduced in order to avoid the bias caused in the case of ApEN [26], resulting from the counting of self-matches [27, 28]. Samp-En indicates more self-similarity in signal analysis and it is a simpler algorithm, compared to ApEn and, that needs approximately half the computational effort. In addition, via the CSE algorithm, one can analyse and quantify the asynchrony between two related signals, estimate the probability of similar patterns between these signals without depending on direction [25].
The CSE algorithm estimates the conditional probability of how many times two similar sequences of *m* points matches to *m+1* points with tolerance *d*. Negative natural logarithm of the previous results, gives the CSE. The CSE algorithm follows [25, 29]:

For two normalized sequences x(i) and y(i), 1≤i≤N, the vector sequences X_i^m and Y_j^m were formed as follows:

$$X_i^m = \{x(i), x(i+1), ..., x(i+m-1)\} \quad (5)$$

$$Y_j^m = \{y(j), y(j+1), ..., y(j+m-1)\} \quad (6)$$

where $1 \le i, j \le N-m$, N is the number of data points of each time series and *m* (embedding dimension) and *r* (tolerance limits of similarity) are fix parameters.

The distance between X_i^m and Y_j^m is defined as:

$$d_{i,j}^m = d[X_i^m, Y_j^m] = \max\left|x(i+k) - y(j+k)\right| \quad (7)$$

where $1 \le k \le m-1$.

For each $i \le N-m$, denote:

$$B_i^m(r)(x \| y) = \frac{number\_of\_j\_that\_meets\_d_{i,j}^m \le r}{N-m} \quad (8)$$

and

$$A_i^m(r)(x \| y) = \frac{number\_of\_j\_that\_meets\_d_{i,j}^{m+1} \le r}{N-m} \quad (9)$$

CSE is defined as:

$$Cross-SampEn(m,r,N) = -\ln\left(\frac{\sum_{i=1}^{N-m} A_i^m(r)(x \| y)}{\sum_{i=1}^{N-m} B_i^m(r)(x \| y)}\right) \quad (10)$$

High asynchrony corresponds to high CSE values [25].

III. PROPOSED METHODOLOGY

The overall feature selection framework is shown in the flow chart of [Fig. **2**].

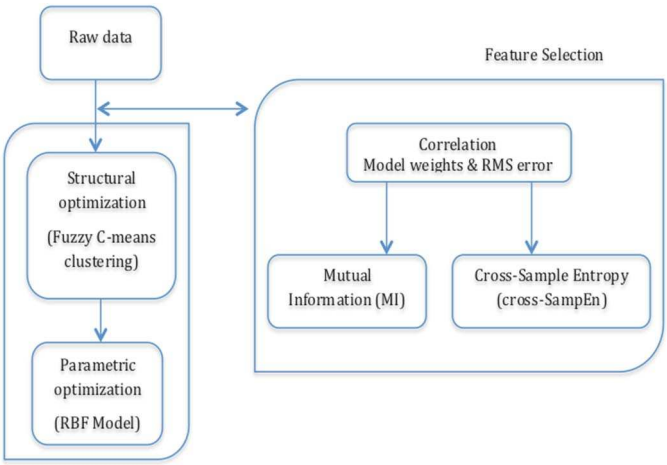

Fig. 2 Feature selection framework - flow chart

The process starts by a data pre-processing step that includes the normalisation of the data set. The dataset is also split into three separate sub-sets, a training dataset (to train the model), a checking dataset (to check for over-fitting during training)

and finally a testing dataset (for testing the model's performance after training).

A. Structural and Parametric Optimisation

The optimisation of the model is performed in two steps, the structural optimisation of the RBF-NF model (in terms on number of rules), and the parametric optimisation of the model's weights (centre and sigma for each membership function, as well as output TSK weights) [19]. Fuzzy c-means is used to cluster the raw data. This produces the desired number of rules (via heuristic adjustment as well as use of cluster validity measures), which also includes the initial values of centre and sigma for each of the membership functions. The parametric optimisation of the modelling structure follows, which includes the use of an adaptive-EP algorithm. Again, heuristically, the best optimisation parameters need to be established, such as the total number of training epochs, the initial rates of learning and momentum for the gradient descent, as well as the decreasing and increasing factor for the adaptive weights [19].

B. Feature Selection

The novelty of the proposed methodology is in the use of the correlation between the output layer weights (TSK) [Fig.3] of the RBF model to the model's training RMSE [Fig.4] as a measure of input (feature) relevance.

Fig.3 shows an example of how the feature weights in the TSK output polynomial change during training (for 300 iterations) for one rule. The TSK polynomial for each rule is of the form:

$$Zi = w_1 x_1 + w_2 x_2 + ... + w_j x_j + ... + w_m x_m \qquad (11)$$

where w_j is the weight for the correspond input x_j.

Similar to regression analysis, the goodness of fit can be correlated to the coefficient of determination to estimate (and rank) the relevance of a particular feature in a dataset. In this paper an information theoretic approach is used, and the goodness of fit (RMSE in this case) is correlated to the variance of individual feature weights in the TSK polynomial.

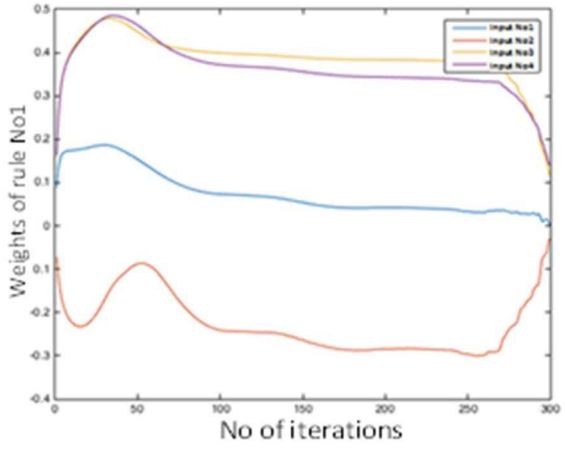

Fig. 3 TSK output layer example: feature weights per rule

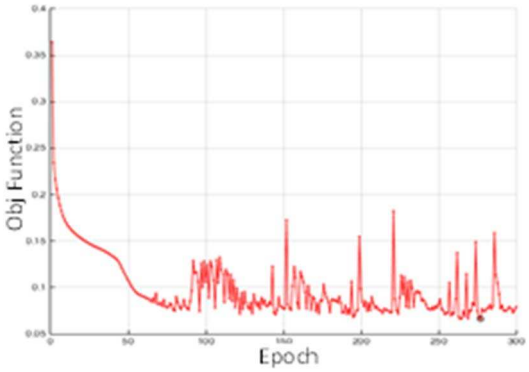

Fig. 4 Model training performance - RMSE

Two measures are tested for their effectiveness in the feature selection task, MI and CSE. Both implementations provide a numerical index that presents the relationship between the two vectors, in our case the two vectors include the feature's weight vector in the TSK polynomial and the model's training performance measured as the RMSE. The proposed approach is calculated for each rule, and repeated for the rest of the Fuzzy Logic rules in the rulebase. Final feature relevance is derived, after aggregating the relevance of each feature across all Fuzzy Logic rules. In the case of MI, higher indices' values correspond to more important features [3], and for CSE the inverse can be assumed [25] (i.e. lower absolute value). This information can then be used to rank all the features in the dataset in terms of their relevance/importance to the model's predictive performance.

IV. SIMULATION RESULTS

To demonstrate the effectiveness of the proposed methodology, three case studies of 4, 5 and 15 dimensional spaces are reported here. First we explore the proposed method's ability to rank relevant features by using the **Iris plant dataset** [30] that is one of the most used literature case study. The second case study under simulation is the **Airfoil Self-Noise** [30] problem, which represents a manufacturing/design example. The third case is also from the manufacturing sector, and it involves the **Heat Treatment of Steel** [31].

The use of these three data sets secures three main characteristics for the evaluation procedure:

- Reliability in the data sets, that already have been used and tested in previous studies (for the purpose of fair comparison).
- Inclusion of both categorical and continuous data.
- Variety in the number of instances/samples, ranging from 150 to a few thousand samples.

For the evaluation of the proposed method, the three datasets were normalised and then randomised. As discussed in the methodology section, three sub-sets were created (using random sample selection) for each case study. Subsequently, FCM clustering was used for the structural optimisation of each model and adaptive-EP for the parametric optimisation.

As detailed in Section III, the parameters for the structural and parametric optimisations were established heuristically. For the purpose of simulation consistency, and to be able to perform a fair comparison of the methodology between different case studies, in all three cases the datasets were separated in a similar fashion. The data sampling was randomly performed, and resulted in approximately 55% samples for training, 20% for checking and 25% for testing.
For the application of the MI and CSE the process of model training and feature selection was repeated for each dataset after inverting the vector direction. Statistical analysis (t-test) between both directions it was carried out to confirm that the proposed algorithm doesn't depend on data direction (common challenge in Samp-En algorithms). Finally, the above process was repeated a number of times for checking the repeatability of the algorithm/results. The simulation results are presented in the following sub-section, shown 'per case study'.

A. Case Study: IRIS dataset

The Iris dataset contains three main categories, namely; a) Iris Setosa, b) Iris Versicolour and c) Iris Virginica of 50 instances each, where each category refers to a type of an iris plant.
Systematic simulations were carried out to establish that the best model performance is obtained via fifteen (15) Fuzzy Logic rules and three hundred (300) training epochs, representing a good compromise between model accuracy and overall computational simplicity.

TABLE **I**, presents the MAE between the model's prediction and the actual output; the model's classification accuracy is also shown in percentage as mean±sd (%). There is a similar performance level between the training and checking sets, hence ensuring avoidance of over-fitting, and the also good testing performance reveals the good generalisation properties of the RBF-NF modelling structure. Overall, the model exhibits good predictive performance, comparable to existing published work, hence it can be considered as a reliable model for use in feature selection [32, 33].

TABLE I. IRIS - MAE AND MODEL CLASSIFICATION ACCURACY PERCENTAGE (%)

	IRIS - MAE% and Percentage of Classification Accuracy	
	MAE% (mean±sd)	***Classification Accuracy% (mean±sd)***
Training	2.59±0.11	100.00±0.00
Checking	2.23±0.05	97.22±0.68
Testing	3.67±0.98	95.61±1.52

TABLE **II**, presents the order of significance of input parameters in IRIS dataset using both MI and CSE as proposed in this paper.

TABLE II. IRIS – FEATURE SELECTION ACCURACY

Feature	**Order of significance of input**	
	MI	***CSE***
Sepal length (cm) (1st)	3	4
Sepal width (cm) (2nd)	4	3
Petal length (cm) (3rd)	**1**	**1**
Petal width (cm) (4th)	**2**	**2**

As in [32, 33], the petal length and width are identified by the proposed feature selection algorithm as the most important variables for classification in the IRIS dataset, hence both variations of the algorithm successfully identify the important parameters for this case study.

B. Case Study: Airfoil Self-Noise

This example employs the Airfoil Self-Noise, a NASA-created data set that contains five inputs, namely; a) Frequency [Hertzs], b) Angle of attack [degrees], c) Chord length [meters], d) Free-stream velocity [meters per second] and e) Suction side displacement thickness [meters] and one output, the a) Scaled sound pressure level [decibels]. Airfoil Self-Noise includes 1503 instances, that obtained from a series of aerodynamic and acoustic tests on different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack [30, 34]. In this case study, the best model performance was obtained for fifteen (15) Fuzzy Logic rules and two hundred (200) training epochs.

TABLE **III**, presents the MAE percentage of model in mean±sd (%) form that presented by simulating the three randomized Airfoil datasets and the results confirm the good performance of the predictive model with an average testing error less than 4%.

TABLE III. AIRFOIL – MAE PERCENTAGE (%)

	AIRFOIL - MAE% Percentage
	MAE% (mean±sd)
Training	2.59±0.11
Checking	2.23±0.05
Testing	3.67±0.98

TABLE **IV**, presents the order of significance of input parameters in AIRFOIL dataset by using MI and CSE. In this case a differentiation between the results of the two implementations (MI and CSE) is observed. While four out of the five parameters are similarly ranked, the parameter of frequency is ranked as 1st and 5th, by the MI and CSE implementations respectively. The frequency variable (input) is known to have a non-linear effect on the noise level (output), which can be significant depending on the level of the other variables. In this case, the MI-based algorithm appears to be better suited to identify this correlation correctly.

TABLE IV. AIRFOIL – FEATURE SELECTION ACCURACY

Feature	Order of significance of input	
	MI	***CSE***
Frequency (Hz) (1st)	1	5
Angle of attack (degrees) (2nd)	**3**	**2**
Chord length (meters) (3rd)	5	3
Free-stream velocity (meters per second) (4th)	**2**	**1**
Suction side displacement thickness (meters) (5th)	4	4

C. Case Study: Steel Heat Treatment

This case study is used to evaluate the proposed method in a real industrial case study, where very high data measurement noise is expected. The example consists a data set related to Steel Heat Treatment and consists of 3760 measurements [19]. The dataset has 15 inputs (process parameters), and 1 output (Tensile Strength). In this simulation, the best model behaviour is observed for twelve (12) Fuzzy Logic rules and for three hundred (300) training epochs.

TABLE **V**, presents the MAE percentage of predictive performance of the model; demonstrating the very good overall performance of the model, with less then 1.5% error in the testing dataset.

TABLE V. STEEL – MAE PERCENTAGE (%)

	STEEL - MAE% Percentage
	MAE% (mean±sd)
Training	1.25±0.11
Checking	1.48±0.14
Testing	1.43±0.01

TABLE **VI**, presents the order of significance of input parameters in STEEL dataset following the use of both MI and CSE algorithms, as presented in Section III.

TABLE VI. STEEL – FEATURE SELECTION ACCURACY

Feature	Order of significance of input	
	MI	***CSE***
Sample test depth (1st)	5	2
Sample size (2nd)	12	10
Test size (3rd)	13	14
C% (4th)	6	4
Si% (5th)	8	13
Mn% (6th)	14	12
S% (7th)	2	1
Cr% (8th)	9	9
Mo% (9th)	7	6
Ni% (10th)	3	7
Al% (11th)	4	3
V% (12th)	10	11
Hardening Temperature (13th)	11	5
Cooling Medium (14th)	15	15
Tempering Temperature (15th)	1	8

Existing research work [31], as well as experts' knowledge suggest as the main critical feature for this process the variable of Tempering Temperature. This is indeed the 'control' parameter for heat treatments in steel, which helps the operators, establish and control material properties. This is because of the metallurgical effect (on microstructure) that heat treatment has on steel. Other important parameters include C% and S% content, as well as alloying elements Ni% and Al%. In our simulation results, as in the previous case study, the MI-based algorithm implementation provides the more consistent results, correctly identifying the Tempering Temperature as the most critical parameter, but also correctly identifies the importance of the main chemical elements. The CSE-based implementation also correctly identifies some – but not all – of the main chemical elements, it fails however to rank high enough the Tempering Temperature variable.

V. CONCLUSIONS

In this paper, a popular implementation of Neural-Fuzzy systems is used to create and evaluate an embedded method of feature selection. The Radial Basis Function – Neural Network modelling structure is used, as an equivalent implementation of a Neural-Fuzzy system with universal approximation properties. In the output layer of the modelling structure a TSK Fuzzy Logic implementation is realised, based on a linear polynomial function. Based on the TSK layer of the modelling structure, the proposed algorithm establishes a method for assessing the importance of the model's features (inputs) in correlation to the model's performance, based on information measures. Two information measures are evaluated in this article, Mutual Information, and Cross-Sample Entropy. The proposed methods' results in a systematic approach to creating a modelling structure while also performing in parallel a feature selection task. Using the information measures in the proposed work, one can rank the features of a case study/dataset in terms of their importance to the process. Existing work, demonstrates the use of such information measures for feature selection as univariate filters, as well as the integration of information measures to Neural-Fuzzy modelling for feature selection as wrappers. However, the presented work addresses for the first time the use of information measures within a RBF-NF modelling structure as an embedded method.

The proposed methodology is tested against two popular datasets in the literature (IRIS – plant data, AirFoil – manufacturing/design data), and one more case study relevant to manufacturing - steel making. Results show that the proposed method creates a) accurate models, that can be used reliably for embedded feature selection and b) the feature selection task is performed satisfactorily in all cases, with the Mutual Information –based implementation of the algorithm having better success rates, compared to the Cross-Sample Entropy –based implementation.

Recommendations for further work in this research direction include the wider evaluation of the proposed methodology against other popular feature selection algorithms, as well as the inclusion in the evaluations of more benchmark functions

as well as well real complex case studies that include uncertainty in the data.

ACKNOWLEDGMENT

This work is supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 636902, Topic: Factories of the Future: Process optimisation of manufacturing assets. We would also like to thank TATA Steel, Yorkshire UK, for the provision of the steel heat treatment manufacturing case study.

REFERENCES

1. A.L. Blum, and P. Langley, "Selection of relevant features and examples in machine learning," Artificial Intelligence, vol. 97 (1-2), pp. 245–271, December 1997.
2. S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in Proc. of the 18th International Conference on Machine Learning (ICML '01), San Francisco, CA, USA, vol. 1, pp. 74–81, 2001.
3. Rüdiger Brause, "Real-valued Feature Selection by Mutual Information of Order 2," in Proc. of the 21st IEEE International Conference on Tools with Artificial Intelligence, pp. 597-604, November 2009.
4. S.M. Ryan, A.L. Goldberger, S.M. Pincus, J. Mietus, and L.A. Lipsitz. "Gender- and age-related differences in heart rate dynamics: are women more complex than men?" J. Am .Coll. Cardiol., vol. 24(7), pp.1700–1707, December 1994.
5. F. Roelfsema , S. Pincus , and J. Veldhuis , "Patients with Cushing's Disease Secrete Adrenocorticotropin and Cortisol Jointly More Asynchronously than Healthy Subjects" J. Clin. Endocrinol. Metab. vol. 83(2), pp. 688-692, February 1998.
6. C.E. Shannon, "A mathematical theory of communication", Bell System Technical Journal, vol. 27(3), pp.379-423, July 1948.
7. E. Freyhult, V. Moulton, and P. Gardner, "Predicting RNA structure using mutual information," Appl. Bioinformatics, vol. 4(1), pp. 53-59, 2005.
8. X. Zhou, X. Wang, E.R. Dougherty, D. Russ, and E. Suh, "Gene Clustering Based on Clusterwide Mutual Information," J. Comput. Biol., vol. 11(1), pp. 147-161, July 2004.
9. H. Peng, Fuhui Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," IEEE Transactions on pattern analysis and machine intelligence, vol. 27(8), August 2005.
10. N.A. Eiseman, M.B. Westover, J.E. Mietus, R.J. Thomas, and T.M. Bianchi, "Classification algorithms for predicting sleepiness and sleep apnea severity," J. Sleep Res., vol. 21(1), pp. 101–112, February 2012.
11. U.R. Acharya, F. Molinari, S.V. Sree, S. Chattopadhyay, Kwan-Hoong Ng, and J.S. Suri, "Automated diagnosis of epileptic EEG using entropies," Biomedical Signal Processing and Control, vol. 7(4), pp. 401–408, July 2012.
12. Huawen Liu, Jigui Sun, Lei Liu, and Huijie Zhang, "Feature selection with dynamic mutual information," Pattern Recognition, vol. 42(7), pp. 1330-1339, July 2009.
13. N. Hoquea, D.K. Bhattacharyya, and J.K. Kalitab, "MIFS-ND: A Mutual Information-based Feature Selection Method," Expert Systems with Applications, vol. 41(14), pp. 6371-6385, October 2014.
14. P. Mitra, C.A. Murthy, and K.P. Sankar, "Unsupervised feature selection using feature similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24(3), pp. 301-312, March 2002.
15. Hahn-Ming Lee, Chih-Ming Chen, Jyh-Ming Chen, and Yu-Lu Jou, "An efficient fuzzy classifier with feature selection based on fuzzy entropy," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 31(3), pp. 426-432, June 2001.
16. P. Angelov, "Fuzzily Connected Multi-Model Systems Evolving Autonomously From Data Streams," IEEE Trans. On Systems, Man and Cybernetics – part B, Cybernetics, vol. 41(4), pp. 898-910, August 2011.
17. Huawen Liu, Lei Liu, and Huijie Zhang, "Boosting feature selection using information metric for classification," Neurocomputing, vol. 73(1-3), pp. 295–303, December 2009.
18. J.-S.R. Jang, and C.T. Sun, "Neuro-Fuzzy Modeling and Control," In Proc. of the IEEE, vol. 83(3), pp. 378-406, March 1995.
19. G. Panoutsos, and M. Mahfouf, "A Neural-Fuzzy Modelling Framework Based on Granular Computing: Concepts and Applications", Fuzzy Sets and Systems, vol. 161(21), pp. 2808-2830, November 2010.
20. J. Park, and I.W. Sandberg, "Universal Approximation Using Radial-Basis-Function Networks," Neural Computation, vol. 3(2), pp. 246-257, June 1991.
21. C.E. Shannon, and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press, Urbana, IL, 1949.
22. R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Transactions on Neural Networks, vol. 5(4), pp. 537-550, July 1994.
23. F. Fleuret, "Fast binary feature selection with conditional mutual information," J. Mach. Learn. Res., vol. 5, pp. 1531-1555, November 2004.
24. B.M. King, and B. Tidor, "MIST: Maximum Information Spanning Trees for dimension reduction of biological data sets," Bioinformatics, vol. 25(9), pp. 1165–1172, March 2009.
25. J.S. Richman, and J.R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," Am. J. Physiol. Heart Circ. Physiol., vol. 278(6), pp. H2039–H2049, June 2000.
26. S. M. Pincus, "Approximate entropy as a measure of system complexity," Proceedings of the National Academy of Sciences of the United States of America, vol. 88, pp. 2297–2301, March 1991.
27. S.M. Pincus, "Approximate entropy (ApEn) as a complexity measure," Chaos, vol. 5(1), pp. 110–117, March 1995.
28. S.M. Pincus, "Quantifying complexity and regularity of neurobiological systems," Methods Neurosci., vol. 28, pp. 336–363, December 1995.
29. T. Zhang, Z. Yang, and J.H. Coote, "Cross-sample entropy statistic as a measure of complexity and regularity of renal sympathetic nerve activity in the rat," Exp. Physiol., vol. 92(4), pp. 659–669, July 2007.
30. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html last accessed 29/1/16
31. G. Panoutsos, and M. Mahfouf, "Granular computing and evolutionary fuzzy modeling for mechanical properties of alloy steels," In Proc. of the 16th International Federation of Automatic Control World Congress, Prague, Czech Republic, pp. 1712-1712, July 2005.
32. M. Dash, and H. Liu, "Feature Selection for Clustering," In Proc. of the 4th Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 110-121, 2000.
33. J.G. Dy, and C.E. Brodley, "Feature Selection for Unsupervised Learning," Journal of Machine Learning Research, vol. 5, pp. 845-889, August 2004.
34. K. Lau, "A neural networks approach for aerofoil noise prediction," Master thesis, Department of Aeronautics. Imperial College of Science, Technology and Medicine, London, United Kingdom, 2006.